IN THE CLAIMS

Please amend the claims as follows:

1. (Currently Amended) A method, comprising:

electronically capturing visual features associated with a speaker speaking during a training session, wherein the visual features include a face recognition of the speaker and a mouth recognition within pixels associated with the face that detects when the mouth is moving and when the mouth is not moving by differences in the pixels from frame to frame in the captured visual features during the training session;

electronically capturing audio during the training session, and wherein the visual features are separated from the audio and processed separately, and wherein the visual features and audio are initially captured at a different rate from one another and separated based on time when each was captured;

matching selective portions of the audio with the visual features by detecting bands of frequencies in the captured audio during same time slices of the training session for the captured visual features when the mouth is detected as moving and when the mouth is detected as not moving to determine when the speaker is speaking and when the speaker is not speaking during the training session; and

identifying the remaining and unmatched portions of the audio as potential noise not associated with the speaker speaking.

2. (Original) The method of claim 1 further comprising:

electronically capturing additional visual features associated with a different speaker speaking; and

matching some of the remaining portions of the audio from the potential noise with the additional speaker speaking.

3. (Original) The method of claim 1 further comprising generating parameters associated with the matching and the identifying and providing the parameters to a Bayesian Network which models the speaker speaking.

4. (Previously Presented) The method of claim 1 wherein electronically capturing the visual features further includes processing a neural network against electronic video associated with the speaker speaking, wherein the neural network is trained to detect and monitor the face of the speaker.

5. (Previously Presented) The method of claim 4 further comprising filtering the detected face of the speaker to detect movement or lack of movement in the mouth of the speaker.

6. (Cancelled).

7. (Original) The method of claim 1 further comprising suspending the capturing of audio during periods where select ones of the captured visual features indicate that the speaker is not speaking.

8. (Currently Amended) A method, comprising:

monitoring an electronic video of a first speaker and a second speaker during a training session for face recognition of the first and second speakers and for indications as to when mouths for the first and second speakers are moving and not moving from frame to frame of the video during the training session;

concurrently capturing audio associated with the first and second speaker speaking during the training session, the audio separated from the video and matched back to a corresponding portion of the video via a particular time slice associated with both the audio and the video, and wherein the video and the audio are initially captured at a different rate from one another are separated and then compared based on time;

analyzing the video to detect when the first and second speakers are moving their

respective mouths by detecting differences in pixels within the faces occurring from frame to frame of the video for each of the speakers; and

matching portions of the captured audio to the first speaker and other portions to the second speaker based on the analysis by detecting bands of frequencies within the audio for a same time slice that indicates a particular mouth of one of the speakers is moving and by noting a particular band of frequency for a particular one of the speakers to indicate that speaker is speaking, and wherein at least some points in the training session indicate that the first and second speakers are simultaneously speaking and discerning what each is saying based on their respective bands frequencies that were noted.

9. (Original) The method of claim 8 further comprising modeling the analysis for subsequent interactions with the first and second speakers.

10. (Previously Presented) The method of claim 8 wherein analyzing further includes processing a neural network for detecting the faces of the first and second speakers and processing vector classifying algorithms to detect when the first and second speakers' respective mouths are moving or not moving.

11. (Cancelled).

12. (Original) The method of claim 8 further comprising suspending the capturing of audio when the analysis does not detect the mouths moving for the first and second speakers.

13. (Original) The method of claim 8 further comprising identifying selective portions of the captured audio as noise if the selective portions have not been matched to the first speaker or the second speaker.

14. (Original) The method of claim 8 wherein matching further includes identifying time dependencies associated with when selective portions of the electronic video were monitored and when selective portions of the audio were captured.

15. (Currently Amended) A system, comprising:

a camera;

a microphone; and

a processing device, wherein the camera captures video of a speaker and communicates the video to the processing device as a plurality of frames within a period of time designated as a training session and each frame associated with a particular time slice, the microphone captures audio associated with the speaker and an environment of the speaker and communicates the audio to the processing device, and the video and audio separated from one another and captured at different rates from one another, and wherein the audio is also associated with the particular time slice of the training session, the processing device includes instructions that identifies visual features of the video where the speaker is speaking and uses time dependencies to match portions of the audio to those visual features to determine that the speaker is speaking, and wherein the processing device recognizes a face of the speaker in each frame of the video and a mouth within the face and detects when the mouth is moving or not moving from frame to frame of the video by changes in pixels associated with the mouth, and wherein when the mouth is moving a detected band of frequency within the same time slice of the audio identifies the speaker as speaking and a particular band of frequency that uniquely identifies the speaker when the speaker is not speaking.

16. (Original) The system of claim 15 wherein the captured video also includes images of a second speaker and the audio includes sounds associated with the second speaker, and wherein the instructions matches some portions of the audio to the second speaker when some of the visual features indicate the second speaker is speaking.

17. (Previously Presented) The system of claim 15 wherein the instructions interact with a

neural network to detect the face of the speaker from the captured video.

18. (Previously Presented) The system of claim 17 wherein the instructions interact with a pixel vector algorithm to detect when the mouth associated with the face moves or does not move within the captured video.

19. (Original) The system of claim 18 wherein the instructions generate parameter data that configures a Bayesian network which models subsequent interactions with the speaker to determine when the speaker is speaking and to determine appropriate audio to associate with the speaker speaking in the subsequent interactions.

20. (Currently Amended) A machine accessible medium having associated instructions, which when accessed, results in a machine performing:

separating audio and video associated with a speaker speaking during a training session into separate frames for analysis, the audio and video originally captured at a different rate from one another and later associated via time when captured, and wherein each frame associated with a same time line to permit specific frames of the video to be matched to specific bands of frequencies of the audio during a same time slice occurring along the time line for the training session to determine when the speaker is speaking and when the speaker is not speaking;

identifying visual features from the video that indicate a mouth of the speaker is moving or not moving by first identifying a face of the speaker and then identifying pixels within the face that represents the mouth and then noting changes in the pixels from frame to frame of the video along the time line; and

associating portions of the audio with selective ones of the visual features that indicate the mouth is moving by matching other bands of frequencies of the audio with detected movements of the mouth during a same time period within the time line and associating a particular band of frequency with the speaker when the mouth is moving indicating that the speaker is speaking.

21. (Original) The medium of claim 20 further including instructions for associating other portions of the audio with different ones of the visual features that indicate the mouth is not moving.

22. (Original) The medium of claim 20 further including instructions for:

identifying second visual features from the video that indicate a different mouth of another speaker is moving or not moving; and

associating different portions of the audio with selective ones of the second visual features that indicate the different mouth is moving.

23. (Previously Presented) The medium of claim 20 wherein the instructions for identifying further include instructions for:

processing a neural network to detect the face of the speaker; and

processing a vector matching algorithm to detect movements of the mouth of the speaker within the detected face.

24. (Original) The medium of claim 20 wherein the instructions for associating further include instructions for matching same time slices associated with a time that the portions of the audio were captured and the same time during which the selective ones of the visual features were captured within the video.

25. (Currently Amended) An apparatus, residing in a computer-accessible medium, comprising:

face detection logic;

mouth detection logic; and

audio-video matching logic, wherein the face detection logic detects a face of a speaker within a video, the mouth detection logic detects and monitors movement and non-movement of a mouth included within the face of the video, and the audio-video matching logic matches specific frequencies occurring within captured audio with any movements identified by the mouth detection logic during a training session and for a same time slice of that training session

to determine when the speaker is speaking and when the speaker is not speaking, and wherein the mouth is detected as moving by changes in pixels that represent the mouth within the face that occur from frame to frame of the video, and wherein the video and audio are initially captured together and at a different rate from one another and separated for analysis and then re-associated via time when each was captured.

26. (Original) The apparatus of claim 25 wherein the apparatus is used to configure a Bayesian network which models the speaker speaking.

27. (Original) The apparatus of claim 25 wherein the face detection logic comprises a neural network.

28. (Cancelled).